SYNTEZA MOWY W E-LEARNINGU DLA OSÓB NIEPEŁNOSPRAWNYCH

Krzysztof Szklanny
kszklanny@pjwstk.edu.pl
Polsko-Japońska Wyższa Szkoła Technik Komputerowych

Słowa kluczowe: <synteza mowy, interfejs, osoby niepełnosprawne, niedowidzący, niewidomi>

Streszczenie:

Temat zdalnego nauczania dla osób niepełnosprawnych staje się coraz bardziej popularny. Rozwijające się technologie multimedialne pozwalają na coraz większą interakcję z użytkownikiem przez co stają się dostępne dla osób najbardziej potrzebujących. Niestety ilość materiałów edukacyjnych wciąż jest niewielka. Dlatego istnieje potrzeba tworzenia oraz rozwijania nowych innowacyjnych technologii na potrzeby e-learningu.

W niniejszym artykule zostanie opisana technologia syntezy mowy. Przedstawione zostaną również inne technologie, które mogą być wykorzystywane podczas tworzenia aplikacji e-learnigowych dla osób niepełnosprawnych. Należą do nich zagadnienia związane z rozpoznawaniem mowy oraz tworzeniem interaktywnych interfejsów użytkownika oraz portali głosowych. Prezentowane w pracy przykłady powstały w ramach prac badawczych prowadzonych w Polsko-Japońskiej Wyższej Szkole Technik Komputerowych.

Streszczenie w języku angielskim:

The subject of e-learning for disabled people is becoming more and more popular. Developing multimedia technologies allows to obtain higher level of interaction with the user. Unfortunately there is still too little educational support for

disabled people. There is a need of creating and developing new innovative technologies, which could be implemented in polish language for the needs of e-learning. In this article different voice technologies are presented like, speech synthesis speech recognition system for polish language. The implemented voice portal at Polish-Japanese Institute of Information Technology is presented. The ideas of interfaces controlled by detection of the movement are described.

Wprowadzenie

Technologie głosowe są na świecie rozwijane co najmniej od połowy lat 70-tych. Ich główną zaletą jest możliwość stworzenia interakcji między użytkownikiem a komputerem lub umożliwienie interakcji między osobą niewidzącą a komputerem. W artykule zostaną przedstawione opracowane technologie głosowe, które mogą stanowić pomoc dla osób niepełnosprawnych. Należy dodać, iż większość produktów skierowanych do osób niepełnosprawnych jest bardzo kosztowna. W artykule zostały zaprezentowane darmowe i gotowe do implementacji technologie.

System Text-to-speech

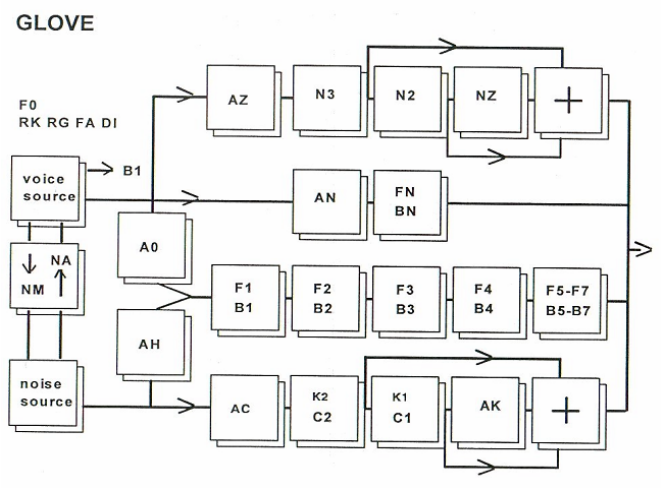
Text-to-speech system czyli moduł konwersji tekstu na mowę jest jedną z metod, która już wielu lat jest stosowana jako pomoc dla osób niewidomych i niedowidzących. Wykorzystuje się tą technologię do generowania dźwiękowej postaci danych tekstowych. Dzięki temu można tworzyć portale głosowe, wirtualne uniwersytety, czy też aplikacje z głosowym interfejsem. Celem nowoczesnych projektów jest zapewnienie takiej jakości syntezy, by słuchający nie był w stanie odróżnić mowy syntetyzowanej od naturalnej. Z oczywistych powodów nie jest możliwe stworzenie i nagranie wszystkich form i wszystkich słów dla danego języka, stąd konieczność syntezowania mowy. System TTS definiuje się jako system

automatycznego generowania mowy z transkrypcją fonetyczną oraz modułami odpowiedzialnymi za prozodię i intonację.

Istnieje kilka metod generowania syntetycznej mowy. Obecnie stosowane są dwie technologie. Pierwsza, zwana regułową syntezą mowy, polega na jej generowaniu poprzez układ symulujący ludzki aparat mowy o zmiennych parametrach. Druga, zwana konkatenacyjną syntezą mowy polega na łączeniu jednostek akustycznych wybieranych z bazy nagrań głosu naturalnego.

Przykładem syntezy regułowej jest syntezator formantowy. Model syntezatora formantowego sprowadza się do zaprojektowania odpowiednich filtrów cyfrowych generujących sygnał harmoniczny bądź szumowy i odpowiednio go modulujących. Jak wiadomo głoski składają się z określonych formantów[1]. Celem syntezatora jest wygenerowanie częstotliwości formantowych. Stąd też nazwa tego rodzaju syntezy. System ten w uproszczony sposób generuje mowę dlatego brzmi ona nienaturalnie i posiada charakterystyczne metaliczne brzmienie. [4, str. 74]

Na Rys. 1. przedstawiony został schemat syntezatora formantowego. Oznaczenie Fx na filtrze, oznacza częstotliwość formantową, gdzie x jest numerem formantu, zaś Bx oznacza kolejne rezonanse filtru.



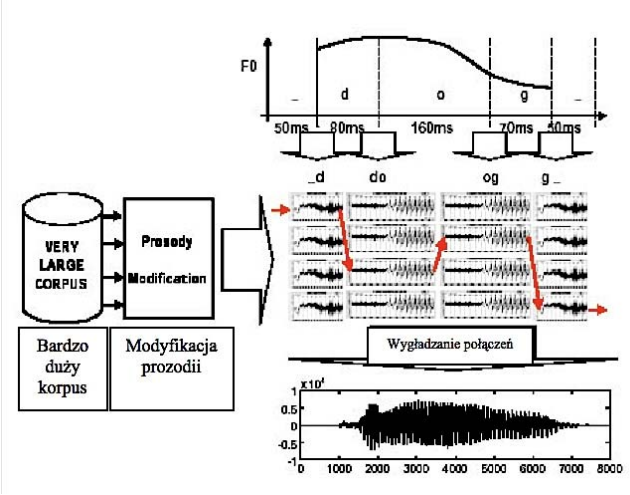
Rys. 1. Formantowy syntezator mowy według Dennisa Klatta. [1]

Drugim rodzajem syntezy jest synteza konkatenacyjna. Główną jej zaletą jest niewielki rozmiar bazy danych, oraz mała ilość jednostek akustycznych. Im mniejszy

1 Formant jest to maksimum w obwiedni widma, tworzonym przez rezonans akustyczny toru głosowego.

rozmiar bazy, tym szybciej będzie syntetyzowana mowa oraz wymagania sprzętowe będą mniejsze. Pewną wadą syntezy konkatenacyjnej jest brak naturalnej intonacji w syntetycznej mowie.[7, str. 2] Rozwiązaniem tego problemu może być użycie syntezy korpusowej (unit-selection). Zamiast tworzenia bazy danych zawierającej tylko jedną instancję danej jednostki syntezy przechowuje się wiele jej realizacji i w zależności od aktualnego kontekstu wybiera się tą najbardziej pasującą. Problemem jest znacznie dłuższy czas generowania zdania oraz sposób wybierania jednostek tak, żeby uzyskać jak najbardziej naturalna mową.[4, str. 76]

Na Rys. 2. znajduje się schemat syntezatora korpusowego.



Rys. 2. Schemat działania syntezy korpusowej.

Najgorszą jakość generowanej mowy pod względem perceptualnym oferuje synteza formantowa. Okazuje się, że z szybkość generowania mowy w syntezie formantowej jest jej jednak bardzo dużą zaletą. Osoby niewidome preferują ten rodzaj syntezy, gdy konieczna jest szybkość obsługi systemu czy też krótki czas dostępu do informacji. Należy dodać, iż w przypadku odsłuchania audiobooków czy też odczytywania napisów do filmów znacznie częściej używana jest synteza korpusowa. Udowodniono, że osoby niewidome są bardzo wrażliwe na sztuczną intonację. W formantowej syntezie mowy nie istnieje moduł intonacji. Generowanie zdań syntetycznych sprowadza się jedynie do zapewnienia podstawowego podobieństwa do generowanych głosek. Kolejną zaletą syntezy jest jej szybkość. W komercyjnych systemach syntezy korpusowej mówi się o rzeczywistym czasie syntezy. W praktyce jednak czas oczekiwania na wygenerowanie, szczególniej pierwszej sekwencji osiąga

wartość nawet kilku sekund. W syntezie formantowej nie istnieje czas oczekiwania na pierwszą sekwencję słów. Dodatkowo możliwa jest zmiana szybkości odtwarzanych zdań. Osoby niewidome lub niedowidzące korzystają z tej funkcji bardzo często. Tak wygenerowane zdanie na syntezatorze formantowym staje się w ogóle niezrozumiałe dla przeciętnego człowieka. Fonetycy twierdzą, iż w niektórych systemach wygenerowana mowa znajduje się na granicy percepcji i zrozumiałości.

Dążenie do uzyskania dużej naturalności generowanej mowy ma swoje szersze zastosowanie wśród osób dobrze widzących. Dlatego też formantowa syntezy mowy do dnia dzisiejszego cieszy się dość dużą popularnością. Rozwój technologii, uzyskanie generowania syntetycznej mowy w czasie rzeczywistym oraz przede wszystkim możliwość modyfikacji czasu trwania przy zachowaniu jego jak największej naturalności z pewnością doprowadzi do popularyzacji systemów typu korpusowego wśród osób niepełnosprawnych.

Synteza korpusowa w systemie Festival

W ramach realizowanej przez autora pracy doktorskiej został stworzony korpusowy syntezator mowy w metasystemie Festival[2]. Festival jest obecnie jedną z najlepiej rozwiniętych platform do realizacji systemów syntezy mowy. Jest pełnym system Text-to-speech z dodatkowymi modułami umożliwiającymi przetwarzanie sygnału mowy. Został napisany w dwóch językach w C/C++ oraz języku Scheme. [6, str. 18]

Festival jest systemem typu Open Source. Jest stosowany w komercyjnych systemach AT&T. Festival może być skompilowany w każdym środowisku unixowym. Istnieje możliwość kompilacji Festivala w systemie Windows. [9, str. 8]

W celu realizacji korpusowej syntezy mowy da języka polskiego autor przygotował:

- moduły lingwistyczne odpowiedzialne za przetwarzanie tekstu,

2 Festival został stworzony na Uniwersytecie w Edynburgu w Centrum Przetwarzania Sygnału Mowy (Centre for Speech Technology Research)

- szeroko pojętą parametryzację sygnału oraz ekstrakcję cech na poziomie akustycznym oraz fonetycznym,
- model akcentów oraz czasu trwania poszczególnych jednostek akustycznych,
- funkcję kosztu, odpowiedzialną za dobór jednostek do syntezy. Autor dla optymalizacji funkcji kosztu zastosował strategię ewolucyjną, co stanowi temat pracy doktorskiej.

Opracowana technologia pozwala na uzyskania dobrze i naturalnie brzmiącej syntezy. Przykłady syntezy zostały umieszczone w prezentacji. Stworzony system może być wykorzystywany do sczytywania informacji tekstowych znajdujących się na stronach internetowych. System jest darmowy i jest dostępny na zasadzie licencji typu „free". [8, str. 2]

Portal głosowy

Kolejną technologią stanowiącą dużą pomoc w e-learningu dla osób niepełnosprawnych są portale głosowe. Na rynku polskim technologia ta jest na etapie rozwoju. Głównym powodem powstrzymującym rozwój portali był brak systemu do rozpoznawania mowy. W niniejszym rozdziale zostanie zaprezentowany system do rozpoznawania mowy oraz portal głosowy zaimplementowany w Polsko-Japońskiej Wyższej Szkole technik Komputerowych.

Informacja głosowa jest bardziej efektywna od informacji tekstowej. Szczególnie, jeśli mówi się o krótkiej informacji: na przykład zalogowanie się do systemu uwaga, komunikat itp. Portale głosowe są tego najlepszym przykładem. Zadaniem portali głosowych jest symulowanie interakcji głosowej z użytkownikiem. Portale głosowe są wyposażone w wyrafinowane mechanizmy interakcji z użytkownikiem, których podstawą jest rozpoznawanie oraz konwersja tekstowej informacji pobranej z bazy danych do postaci dźwiękowej. Portal głosowy jest nie tylko wymyślnym systemem do prowadzenia konwersacji z komputerem, lecz przede wszystkim stanowi bazę danych, czyli zasób ważnych informacji dla potencjalnych

klientów serwisu. Informacje te przechowywane są w postaci tekstowej na serwerach baz danych, skąd pobierane są przez skrypty, zlokalizowane na serwerach WWW, obsługujące zapytania SQL. Wyselekcjonowane wiadomości konwertowane są do postaci dźwiękowej przez przeglądarkę głosową i emitowane. Portal głosowy posiada wbudowany syntezator mowy, który odpowiednio sczytuje informacje z ekranu i zamienia je na postać głosową.

Stworzenie dobrego systemu do rozpoznawania mowy niezależnego od mówcy jest zadaniem nietrywialnym. Do rozpoznawania mowy używa się HMM, sieci neuronowych lub połączeniu obydwu technologii. W Polsko-Japońskiej Wyższej Szkole Technik Komputerowych zaimplementowany został portal głosowy Primespeech[3]. Jest to telefoniczny systemem informacyjny wyposażony w technologie automatycznego rozpoznawania mowy oraz syntezy mowy. Jest to pierwszy taki system stworzony dla języka polskiego. Innowacyjność technologii pozwala w dniu dzisiejszym na implementację i automatyzację systemów telefonicznych oraz tworzenia serwisów sterowanych głosem. Portal umożliwia przełączenie mówcy do 150 różnych pracowników poprzez rozpoznanie imienia i nazwiska bądź wypowiedzianego miejsca. Studenci mają możliwość uzyskania informacji o ocenach, wybranych przedmiotach w indywidualnym toku nauczania, saldzie oraz numerze konta i statusie studenta. Dzięki zaimplementowanej technologii syntezy mowy portal głosowy syntezuje przez telefon najnowsze ogłoszenia dziekanatu i aktualności uczelniane.

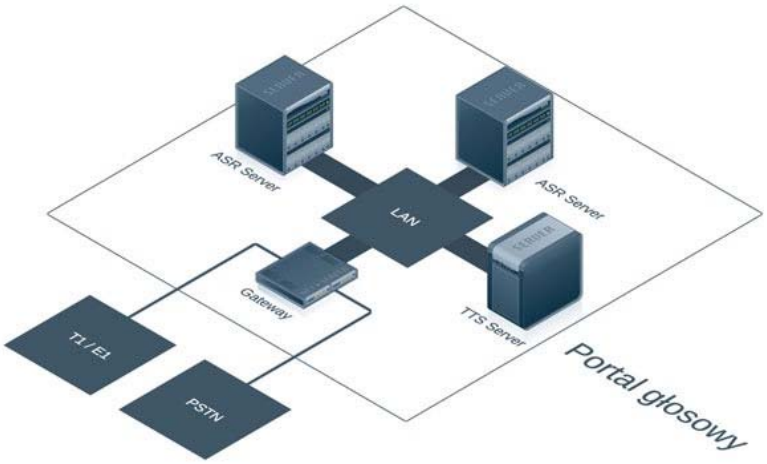
Opracowana technologia jest rozszerzalna i w znacznej mierze może ułatwić osobom niewidzącym korzystanie z treści edukacyjnych dostępnych za pomocą technologii głosowych. Należy dodać, że połączenie obu technologii coraz częściej stosowane jest w e-learningu.

3 www.primespeech.pl

Architektura portalu

Portal głosowy jest rozproszonym systemem sieciowym. Składa się z trzech podstawowych programów, które komunikują się ze sobą przez sieć lokalną: Gateway, TTS Server (Text-To-Speech Server) i ASR Server (Automatic Speech Recognition Server).

Na Rys. 3. przedstawiono architekturę systemu Primespeech.



Rys. 3. Architektura portalu głosowego.

ASR Server stanowi "serce" portalu głosowego. Jego głównym zadaniem jest rozpoznawanie mowy. Został on skonstruowany w oparciu sieci neuronowe. ASR Server posiada wbudowany moduł adaptacji akustycznej, który pozwala na dotrenowanie zespołu sieci neuronowych do nowych głosów. Architektura programu pozwala na uruchomienie dowolnej ilości dialogów na jednym serwerze. Ilość równoległych dialogów jest zależna od prędkości komputera i ilości pamięci.[2, str. 7]

TTS Server to wielowątkowy program ze zintegrowanym syntezatorem mowy. Syntezator mowy może być uruchomiony tylko w kilku instancjach dlatego TTS Server synchronizuje polecenia syntezy tak, aby zostały one wykonane w optymalnej kolejności. Syntetyczna mowa może być opcjonalnie zapisana na dysk, aby uniknąć w przyszłości ponownej procesu syntezy mowy. [11]

Primespeech Voice Portal Manager jest systemem zarządzania portalem głosowym przez dowolną przeglądarkę internetową. Pozwala on na zalogowanie się do serwisu przez bezpieczny protokół SSL. Po zalogowaniu się użytkownik może

modyfikować ustawienia portalu poprzez zmianę ilości operatorów dostępnych w infolinii, zmodyfikowanie bazy osób i telefonów. Możliwe jest zmiana informacji odtwarzanych przez portal. Dodatkowo istnieje możliwość przeglądania statystyk i wykresów dotyczących dystrybucji rozmów w czasie, długości przeprowadzonych rozmów czy procentowego rozkładu pytań w zależności od wybranego tematu.[11]

Ostatnim elementem systemu jest Primespeech Gateway. Jest to program, który działa na komputerze z kartą telefoniczną Dialogic. Używana karta telefoniczne posiada procesor DSP na każdy kanał, co pozwala na rozpoznawanie mowy w czasie rzeczywistym. [3, str. 2]

Przedstawiona technologia portalu głosowego została przetestowana i wdrożona. Istnieje możliwość wdrożenia i udostępnienia systemu osobom niepełnosprawnym i państwowym instytucjom naukowym na zasadzie licencji darmowej, pod warunkiem nie wykorzystywania systemu do działalności komercyjnej.

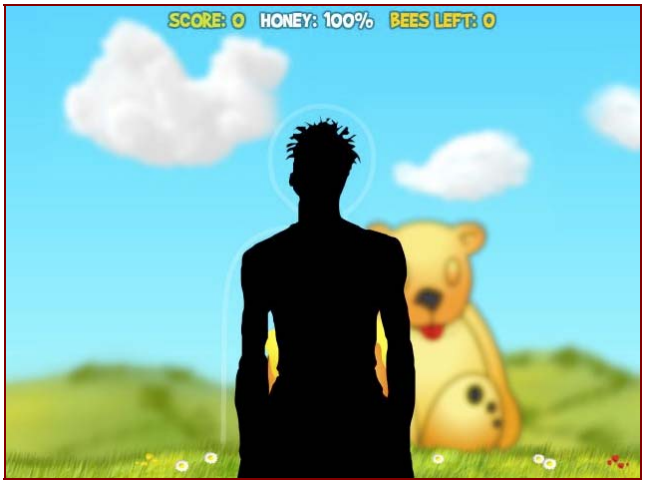
Nowoczesne interfejsy dla osób niepełnosprawnych

Standardowe interfejsy aplikacji stanowią duże utrudnienie dla osób niepełnosprawnych. Szczególnie problematyczne są interfejsy, które nie zostały przetestowane pod względem użyteczności. Interesującą propozycją dla osób niepełnosprawnych jest interfejs, który pozwala na kontrolowanie aplikacji w sposób interaktywny. Przykładem takiej aplikacji jest multimedialny zestaw gier dla najmłodszych stworzonych w ramach pracy inżynierskiej w Polsko-Japońskiej Wyższej szkole Technik Komputerowych[4]. Zamiast myszki stosowana jest kamera internetowa sczytująca pozycję rąk. Aplikacja zawiera kilka gier multimedialnych skierowanych do najmłodszych. Pomysł ten może być zaimplementowany w aplikacjach dla osób niedowidzących.

Rys. 4. przedstawia omawiany interfejs. Kamera, która jest skierowana na osobę, rozpoznaje ruch rąk, następnie informacje o ruchu przesyłane są do aplikacji, gdzie są interpretowane po czym podejmowana jest odpowiednia akcja w grze. Osoba

4 Aplikacja dostępna jest pod adresem. www.funcam.pjwstk.edu.pl

powinna się znajdować w odpowiedniej odległości od kamery tak, żeby znalazła się w zaznaczonym obrysie człowieka. [5, str. 20]



Rys. 4. Interfejs sterowany ruchem rąk.

Przykładem kolejnej technologii wartej uwagi w aplikacjach e–learningowych dla osób niedowidzących jest aplikacja rozpoznająca twarz osoby oraz wyznaczająca jej odległość od monitora. Taka aplikacja powstała w Polsko-Japońskiej Wyższej Szkole Technik Komputerowych pod kierunkiem prof. Krzysztof Maraska. Możliwe staje się dynamiczne zmienianie wielkości czcionki w zależności od odległości od komputera. Dodatkowym atutem aplikacji jest przechowywanie wielu profili osób, automatyczne rozpoznawanie twarzy oraz zdefiniowanie za pomocą RSS informacji, które mają być wyświetlone lub przeczytane w przypadku osoby niewidomej. Rys. 5 przedstawia omawiany interfejs.[6, str. 13]

Rys. 5. Interaktywny interfejs.

Kolejnym pomysłem, który znacznie ułatwi edukację osobom niewidomym i niedowidzącym jest stworzenie wirtualnego obrazu dla osób niewidomych. Drukarki brajlowskie są bardzo drogie. Trwałość wydruku obrazu z takiej drukarki zależy wyłącznie od sposobu korzystania, niemniej jednak obraz taki może ulec szybkiemu uszkodzeniu. Ciekawym i wartym do zaimplementowania pomysłem wydaje się być implementacja aplikacji do wirtualnych obrazów na konsoli do gier Nintendo Wii. Za pomocą kontrolera, który ma wbudowany czujnik ruchu można sczytywać dane o wirtualnym obrazie. Informacja jest przekazywana osobie niewidomej w postaci wibracji. Stopień wibracji jest nośnikiem informacji o barwie. Takie rozwiązanie umożliwi tworzenie wirtualnych obrazów o znacznie szybszym sposobie czytania takich obrazów oraz o większej ilości barw niż w przypadku brajlowskiego wydruku.

Podsumowanie

W artykule zostały opisane i technologie głosowe – syntezy i rozpoznawania mowy polskiej. Mogą one być wykorzystane i zaimplementowane w portalach głosowych, wirtualnych uniwersytetach. Przedstawiono również technologie tworzenia interfejsów dla osób niepełnosprawnych opartych na detekcji ruchu. Zaprezentowane technologie są technologiami darmowymi. Mogą one w znacznym stopniu ułatwić osobom niepełnosprawnym korzystanie z aplikacji e-learningowych. Autor ma nadzieję, że w niedługim czasie technologie te staną się bardziej popularne i dostępne dla osób niepełnosprawnych.

Literatura

[1] R. Gubrynowicz R. „Podstawy Fonetyki Akustycznej”, Wykład, PJWSTK 2006

[2] D. Koržinek “Hybrydowy System Automatycznego Rozpoznawania Mowy w Języku Polskim”, praca magisterska, PJWSTK 2007

[3] D. Koržinek, Ł. Brocki, R. Gubrynowicz, K. Marasek „Wizard of Oz Experiment for a Telephony-Based City Transport Dialog System”, IIS 2008

[4] K. Szklanny "Przygotowanie bazy difonów dla realizacji syntezy mowy w systemie MBROLA-a."praca magisterska PJWSTK 2002

[5] A.Drzewicki „Fun Cam Gry komputerowe dla dzieci z interfejsem sterowania opartym o detekcję ruchu", praca inżynierska, PJWSTK 2006

[6] M. Zalewski, A. Dąmbski, M. Boksa, M. Więsyk „Interaktywna prezentacja multimedialna wykorzystująca system wizji maszynowej OpenCV" praca inżynierska, PJWSTK 2005

[7] K. Szklanny "Preparing the Polish diphone database for speech synthesis in MBROLA". 50 Otwarte Seminarium z Akustyki Szczyrk, Poland, 2003

[8] D. Oliver, K. Szklanny "Creation and analysis of a Polish speech database for use in unit selection synthesis", LREC Genoa, Italy 2006

[9] D. Oliver "Polish Text to Speech synthesis system", MSc Thesis, Edinburgh University, Edinburgh, 1998

[10] A. Black, P. Taylor 1998. "Festival Speech Synthesis System: system documentation". Technical Report HCRC/TR-83, University of Edinburgh, Human Communication Research Centre.

[11] D. Koržinek , Ł. Brocki, www.primespeech.pl, 2008